Version 2.0
DØ Note 4262

Data Streaming Review Committee Interim Report

Committee: Laurent Duflot, Leslie Groer, Marumi Kado, Avto Kharchilava, Greg Landsberg, Peter Mättig, Rick Van Kooten (chair)
12 September 2003

1. Introduction

Given the high level of activity preparing physics analyses for presentation at Lepton Photon 2003, the desire to use the experiences of data access during this time, and lessons learned from preparations for the p14 reprocessing, it was decided to produce a brief interim summary report. Subsequent information, including first collaborations with the newly created Common Sample Working Group will be included in the final report. Initial recommendations will be presented here and a proposal for an alternative to online streaming outlined.

The charge to the committee is included in an appendix at the end of the report.

2. Background

There is a great deal of documentation and reports concerning streaming that already exist. An effort has been made to collate some of these on the committee's web page at **www-d0.fnal.gov/phys_id/bid/d0_private/streaming_review.html**

The work of the committee proceeded through meetings of the committee, inviting experts to give details on the current streaming tools and plans in addition to feedback on committee proposals, and open meetings where physics group representatives were invited to give input and feedback to the committee. Agendas, files of talk transparencies, and minutes of these meetings can be found on the DØ agenda server: Computing and Software -> Computing and Software Operations for much more detail. Individual talks are also linked from the committee web page at the above address. A status report was presented at the DØ Collaboration Meeting in Beaune, France in June 2003.

3. Benefits of Streaming

Without going into detail, the benefits of streaming are clear, and repeating from the charge:

- guarantees that an analysis based on a specific trigger and/or physics object will not have to process every data file in order to gain access to the events of interests;
- ensure that an analysis will not have to, in effect, access every tape in order to gain access to files of interests;

- minimize the wait time for data to be made available for a particular analysis or processing activity and to minimize the number of times that tapes are mounted and data is read from tape.

For reprocessing, it would allow us to reprocess only "important" subsets of the data, reprocess "more important" data first, and/or reprocess "differently important" data elsewhere[1].

If done online, it helps with online efficiency and offers the flexibility of processing "less important" data later if the offline farm cannot keep up with the online data rate. Offline, it would improve the efficiency of the pick_event tool.

4. Online Streaming

Online streaming can be either inclusive or exclusive streaming. Almost all of the effort in the past 16 months dedicated towards implementing exclusive online streaming.

4.1 Inclusive Online Streaming

The L3 filter code assumes and is designed for exclusive streaming except for the "mark & pass" events that are handled separately and in a special way. For large stream overlaps, online data transfer would have to deal with an additional burden as well as an increase in processing time for events on the farms. A redesign of L3 would involve significant effort.

The luminosity system currently cannot handle duplicate events at the online level and a redesign is predicted to involve significant effort.

For SAM, exclusive streaming was a design constraint for the event catalog, and tools such as pick_events and a possible future implementation of direct access to data would be difficult for inclusive streaming. Some loss in SAM functionality would result and it is difficult to assess all of the consequences and level of effort needed for this redesign.

Online streaming implies duplicate events that result in increased online bandwidth as well as increasing needed tape resources.

A rough estimate of the effort to implement inclusive streaming is 3 FTE-years.

Recommendation 1: the level of effort required to implement inclusive streaming online is too great to commit resources to it at this time.

4.2 Exclusive Online Streaming

In exclusive streaming, each event goes into one and only one stream. Streaming criteria is based on L1/L2 triggers and L3 filter information. It is estimated that about 1.5–2.0 FTE-years has already gone in to effort to implement exclusive online streaming.

[1] A nice one-line summary from Adam Lyon.

4.2.1 Exclusive Streaming Scenarios

Scenarios that were considered for exclusive online streaming:

- Physics priority based: particular event topologies have higher priorities than others and events are diverted according to priority. Those analyzing the higher-priority streams would have to look at the fewest multiple streams.
- Physics object based: e.g., individual streams for single high p_T electron, single high-p_T muon only, high-p_T electron plus high-p_T muon, two muons, two electrons, etc. Could result in many streams, some of which could be small and may require grouping with others. Everybody would have to analyze more than one stream.
- Load balancing based: trigger patterns are analyzed to construct *n* streams of roughly equal size – would include some physics priorities.
- Round robin: stream to *n* equal size streams without selection. Improves online performance, and could selectively process fewer streams if falling behind.

4.2.2 Current Status

The extensive work, particularly by Adam Lyon, that has already been done for the components needed to implement exclusive online streaming is recognized.

L3 is ready to do exclusive streaming at the L3 object level. Using an extensively tested tree algorithm, priority-based streaming would be easy to implement, and physics-object based streaming could be made to work. A tool, evaluateStreaming, exists to study streaming scenarios. A lost file catcher is almost complete. The offline streaming database is ready. Work is needed for dataset creation in SAM for streaming as is modifications to luminosity tools. Work is needed to write tools to conveniently create SAM datasets from streamed data. Furthermore, changes to the online luminosity system are needed for the offline luminosity tools.

It is estimated that approximately 1 FTE-year is needed to completely finish and test the implementation.

4.2.3 Evaluation

Exclusive online streaming clearly provides all of the outlined benefits of streaming. Concerns from the committee and physics groups included relying on L3 objects to make the choices between physical streams, particularly when trigger lists and L3 physics objects change (or the trigger is unstable). At L3 the detector may not be calibrated; an example being a hot cell found online that could be removed at d0reco level. D0reco physics objects that are currently used for skimming and data analyses are currently better understood than L3 physics objects. This situation may clearly change as L3 matures.

For the first two streaming scenarios, physics priorities have to be set early in the processing. Luminosity bookkeeping with exclusive online streams is difficult and would need substantial work. Finally, the additional complexity of doing analyses having to use many streams (e.g., more than 2 or 3) is also a concern.

The opinion of the committee is that there exists a viable alternative of streaming offline, at the output of d0reco, and that it has many of the same benefits of exclusive online streaming. D0reco physics objects are available in addition to trigger information that can be used to form streaming criteria. Flexibility in assigning stream criteria is increased with more available information, most notably offline tracking.

Recommendation 2: Online exclusive streaming is not recommended at this time.

It should be noted that online exclusive streaming still provides several benefits not offered by offline streaming, i.e., streaming online streams data tier RAW events (instead of data tier DST, see next section), and online exclusive streaming allows for selective processing of these RAW event streams in case the farm cannot keep up with the online data rate and improves online efficiency. Since pick_events operates from RAW events, pick_event efficiency would improve. The final set of recommendations addresses these issues.

5. Alternative: Offline Streaming

D0reco on the farm takes RAW events as input from online and outputs DST and tmb events, usually written to tape. At this point, full d0reco reconstructed physics objects are available in addition to trigger and L3 filter information. This information can be used in streaming criteria to decide to which physical offline stream to write the event. In this case, it is data tier DST (and tmb) events that are streamed rather than RAW events.

5.1 Reprocessing from DST

To gain one of the major benefits of streaming, i.e., to be more flexible with reprocessing, it will be necessary in the future to reprocess DST streams if offline streams are implemented. In some cases, reprocessing at the level of clusters and hits cannot be done from DST – one would have to go back to RAW events. Some re-alignment would still be possible. What is available in the DST for reprocessing:

- CFT: ADC values for all hit fibers stored, can redo CFT clusters, alignment, everything;
- SMT: full 1D reconstruction and clusters stored, but not strips, cannot redo clusters, but can redo alignment;
- CAL/PS: clusters stored, CalDataChunk is kept, could redo calibration of ADC to energy but not drop the threshold cut;
- Muons: as of p14, muon hits are available, can redo t0's, remake segments; no framework for re-alignment. Cannot redo hit finding.

It should be noted that part of the p14 reprocessing would already be done from DST. To be able to more quickly answer possible differences between reprocessing from DST

versus RAW, a very small fraction of streamed events (unbiased plus special samples such as $Z \to e^+e^-$) could include the RAW data chunk along with the DST.

5.2 Event Size and Processing Speed

In the evaluation of exclusive versus inclusive offline streaming (i.e., allowing event duplication), event size needs to be taken into account. For p14.04, the average RAW event size is ~280 kB, and the average DST event size is ~176 kB (a reduction from a size of 223 kB in p14.03). These can be compared to sizes used in models of computing resources where a RAW event size of 250 kB and DST event size of 150 kB are assumed. A significantly large fraction of duplicated DST events can therefore substantially strain the DØ tape budget, although clear gains in data access, storage, and data transfer can be made with further reduction in DST size.

The average size of a tmb event is 15–20 kB, i.e., roughly a factor of 10 smaller than DST events. Multiple copies of tmb events are far less serious, and tmb event duplication in physics group skims is routine and planned for in computing resource models.

Processing time estimates are 20–30 GHz-sec/event and the farm represents roughly a THz of computing power with imminent contributions from remote sites. For reasonable event rates, a target to reprocess a year's worth of data within 3–6 months may require selective reprocessing of streams.

Hence the obvious, in addition to considering streaming, is that additional effort should be spent to further reduce the DST size and speed up d0reco.

5.3 Requirements

Code needs to be developed to implement stream selection at the output of d0reco and to write these events to different physical locations. Work continues on the merging of multiple streams separately and SAM dataset definitions of multiple streams. Luminosity tools need to be modified, but the effort and bookkeeping required are modest when streaming is offline. As will be described in more detail later, data structures should be added to DST's and tmb's to store "stream/skim bits" (and version) to record stream and substream decisions.

5.4 Evaluation

Pros for offline exclusive streaming are:

- requires less effort to implement than online streaming;
- has the capability to include the best possible d0reco physics objects to decide stream selection;
- is flexible and can be easily tuned for more optimal data access;
- the bookkeeping for luminosity determination offline can be easier since the reference to the unique RAW event still exists;
- matches the needs for most reprocessing;
- can include the writing of tmb streams closely matching the tmb skims of physics groups reducing this tedious task when done centrally;
- certain subsets of data can be checked for quality quickly using reco_cert;

- existing streaming tools such as evaluateStreaming and the missed file catcher are still useful.

Cons for offline streaming are given at the end of section 4.2.3.

Recommendation 3: Pursue the viable alternative of offline streaming with d0reco.

6. Possible Offline Streaming Scenario

A possible offline streaming scenario is described that could proceed in phases. It takes advantage of work done by the Common Skim Working Group that consisted of representatives from the top, Higgs, NP, and Electroweak Physics Groups to determine common skim requirements that could be shared by the groups.

As of June/July 2003, the criteria going in to the "common skim" were:

Category	*Criteria*	*% of Data*
Single Muon	1 "loose" muon, max[p_T(central),p_T(local) > 8 GeV	~8%
Dimuons	2 "loose" muons, no p_T cuts; or 1 "loose" muon + 1 MTC muon with a matching central track with $p_T > 8$ GeV	~1.5%
Single EM	EM(10,11), $p_T > 17$–18 GeV	~10%
DiEM	2 EM(10,11), $p_T > 7$ GeV	~2%
EM-Muon	1 "loose" muon and one EM(10,11) with $p_T > 5$ GeV	~2%
Electron w/ track	EM(10,1) w/ $p_T > 12$ & a track w/ $p_T > 7$ GeV within $\Delta\Box = 0.4$	~5%
All Jets	3 jets, with $\lvert\Box\rvert < 2.5$ and $p_T^1 > 20$ GeV, $p_T^{2,3} > 15$ GeV	~10%
Jets + MET	MET > 20 GeV, HT > 25 GeV	~8%
Total	Inclusive (w/ overlaps between categories) Exclusive (add, removing overlaps)	**~47%** **~35%**

A reasonable extension of the criteria for the Dimuon category would satisfy the additional needs of the *B* physics group and only increase the size of the category to less than ~4% of the data. Further development of criteria will fall under the auspices of the new Common Sample working group for further optimizing the exact criteria, but the scenario described is changed little as long as the exclusive size of the common skim does not increase too much. Object identification criteria should be as loose as possible to take into account future improvements.

6.1 Phase 1

As in the "strawman" presented at Beaune, at the output of d0reco, exclusively stream offline any DST events that satisfy any of the common skim category requirements (or different future criteria) to the physical location "Stream A". Event category requirements that were satisfied by the event would be flagged with "streaming/skim bits", one per category. These bits should be included in the DST and tmb in such a way that they can later be read without unpacking the entire event.

The common skim criteria do not cover all the physics or object identification needs of DØ, e.g., low-p_T leptons for some important *B* physics analyses and other jet-only topologies for some QCD analyses. As an example, for single muons, dropping to $p_T(\square)$ > 2 GeV gives a stream of ~30–35% of the data. Such low-p_T lepton events that do not satisfy the common skim requirements would then be exclusively written to "Stream B", and the remainder to "Stream C". The result would be three broadly defined exclusive streams, each roughly of equal size. The stream/skim bits would also record this division.

To ease the evaluation of efficiencies and backgrounds of new physics objects as well as the determination of migration between streams when they are later reprocessed, a small unbiased, prescaled fraction of Stream B would be exclusively streamed out as Stream B' (by, for example, writing every fiftieth Stream B event to B') and a similar unbiased, prescaled fraction of Stream C to Stream C'.

These streams would be exclusive and priority-based. For example, a small fraction of low-p_T lepton events will be in Stream A if they satisfy the requirements of the common skim, but they will be identified as also falling in to the Stream B category by the stream/skim bit for later access.

The use of logical streams in SAM (that have yet to be used) allows the easy combining of exclusive streams, e.g. "completeStreamC" could then be defined as the combination of the exclusive Stream C + StreamC'.

Summarizing, exclusive streams:

- Stream A : common skim/sample, ~30–35% of data
- Stream B : low-p_T leptons, ~30–35% of data
- Stream B': small, unbiased, prescaled fraction from Stream B
- Stream C, remainder, other QCD, ~30–35% of data
- Stream C', small, unbiased, prescaled fraction from Stream C

Each of these streams would automatically produce a corresponding tmb stream. Initially, take the Stream A tmb's and on CAB for example, run the production centrally (and automated) to make the *inclusive* tmb skims corresponding to the common skim/sample categories of the physics groups. Once tape drive allocation, merging, and resources have been tweaked, optimized, and deemed to be reliable, move this task to d0reco and the farm.

Even at this point, the DST/tmb size to make the inclusive thumbnails is approximately three times smaller and data access efficiency is improved with each stream grouped on the same tape families. The majority of data analyzers could concentrate on Stream A and the common skim events. It should be stressed that Stream B and Stream C are not

"junk" for anyone – they are also needed to evaluate backgrounds, but the goal is that the smaller Stream B' and C' could satisfy these needs in some cases.

Recommendation 4: Implement as soon as feasible offline exclusive streaming along the lines of "Phase 1" described above.

In the future, it may be possible to migrate some of the offline exclusive streaming to online once L3 is more mature and the D0 trigger capabilities are complete. Such a migration would give us the advantages of online exclusive streaming (see Sec. 4.2.3) and would be facilitated by the work already invested in that system and the experience to be gained with offline streaming.

Recommendation 5: Online exclusive streaming should not be discounted for use in the future.

6.2 Phase 2

Once the Phase 1 level of streaming has been proven to be stable and effective, further division of Stream A (or other streams) at the DST level could be pursued with three possible options of direct access to data, further exclusive streams roughly corresponding to the common skim/sample definitions, or further inclusive offline streams, directly corresponding to the common skim/sample definitions. These options are less definite than Phase 1 and will continue to be fleshed out.

6.2.1 Direct Access to Data

The header of each file already has an index that allows random access to a given collision id/run/event inside the file. This header is valid for i) evpack format events (essentially all our data); ii) if there was not a crash before the end of the file since the header is written last. This functionality is already coded in to ReadEvent and WriteEvent.

In Run 1, d0dad (direct access to data) allowed random access to miniDST's resident on disk. Given a list of run/event/LBN numbers, a catalog would be built giving the physical location of each run/event on disk. It was estimated to take 6–8 weeks of work by someone reasonably skilled with databases to implement a Run 2 analog given the functionality described above already present.

An interesting possibility is if all of the DST, or at least the Stream A DST, could be pinned to disk. Selective reprocessing or remaking tmb's could be done with run/event lists determined using the tmb's. A possibility to explore further is an interface of d0dad with SAM with a preprocessing phase that would find which files had the events satisfying a stream/skim bit mask and then launching a project to get the files.

Having all of the DST's pinned to disk in the current model exceeds current storage projections of 75 TB per year. However, if it was just the common skim Stream A it would require ~ 26 TB/year. Actual disk costs are not unreasonable, but the infrastructure, e.g., space to put them, cooling, networking, etc., are more difficult.

Direct access to data would essentially be a "virtual" and very flexible inclusive streaming scheme.

6.2.2 Further Exclusive Offline Streams

Having further exclusive streams, e.g., as substreams within Stream A, is only a modest extension. The problem will be is if it is possible to have useful exclusive definitions that could roughly correspond to the common skim/sample criteria.

6.2.3 Further Inclusive Offline Streams

The exclusive DST Stream A could be divided further into inclusive DST substreams. This was in fact part of the "strawman" streaming proposal presented at Beaune. In the proposed streaming scenario, the eight individual, inclusive streams of the common skim result in 33% of the events being duplicated. Neglecting correlations of duplication with category event sizes, this would increase the Stream A DST event space required by a factor of ~1.33. The total DST size for all of the data increases by a factor of ~1.12.

Difficulties with offline inclusive streaming were subsequently clarified:

- It would be problematic to combine individual DST streams because of the overlap or if one wants to spin over all the DST's. If one reprocesses an individual inclusive stream (and doesn't "re-stream"), keeping track of migration of events between streams could be complicated bookkeeping. The efficiency for entering the stream could go up (usually), but could also go down. Reprocessing and then re-streaming is only feasible if one spins over all DST's (or at least all of Stream A), which is again complicated to handle event duplication at the reco stage.

- There are differing opinions on the difficulty to obtain precision luminosity if the one has to look at more than one of the inclusive DST streams (e.g., for backgrounds). tmb's are made from inclusive DST's resulting in overlap problems tracing the exact parentage from the tmb's back to the unique raw data file names.

- If all eight Stream A categories were streamed to inclusive DST's and inclusive tmb's, there is a concern regarding the number of tape drives simultaneously needed by d0reco on the farm as well as resources needed for merging all of these files on the fly (in to ~1 GB chunks or for entire runs, whichever comes first). d0reco on the farm has priority over all other tape drive requests and streaming both DST and tmb to tape could cause problems with drive availability.

7. Acknowledgements:

We would like to acknowledge the contributions of colleagues who provided feedback from the physics groups. We also appreciate the valuable expertise and advice from "technical consultants": Amber Boehnlein, Adam Lyon, Wyatt Merritt, and Heidi Schellman.

Appendix 1: Charge to the Committee

Three years ago, an internal DZero review committee proposed to implement exclusive streaming online as a means to

- guarantee that an analysis based on a specific trigger will not have to process every data file in order to gain access to the events of interests;
- ensure that an analysis will not have to, in effect, access every tape in order to gain access to files of interests;
- minimize the wait time for data to be made available for a particular analysis or processing activity and to minimize the number of times that tapes are mounted and data is read from tape.

Given that the exclusive streaming has not been deployed and that we have much more information about our data processing and physics analyses, it is prudent that we review the exclusive streaming model to revisit the above issues and to address the following additional questions:

- In the event that our offline farm cannot keep up with the online data taking, can the exclusive streaming facilitate data processing? What are the impacts of different streams processed with different RECO versions?
- It is unlikely that we will be able to reprocess all events. Can the exclusive streaming model facilitate reprocessing? Minimize computing resources? How does the reprocessing impact on different physics analyses?
- Given what we know now how users access data and how analyses are performed, is the exclusive streaming still the right model?
    - If so, what critical user tools we must have? What are possible streaming scenarios for the current and the expected future trigger lists?
    - If not, what are possible alternatives? How do these alternatives compare with the current exclusive streaming plan in terms of data processing, reprocessing and user analyses? What we need to do to implement those alternatives? What human and computing resources are needed?

Appendix 2: Figures

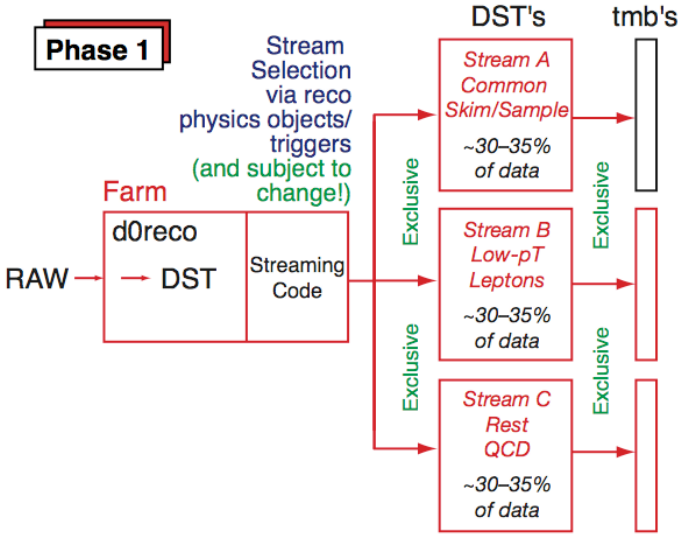

Figure 1: Schematic of proposed "Phase 1" of exclusive offline streaming.

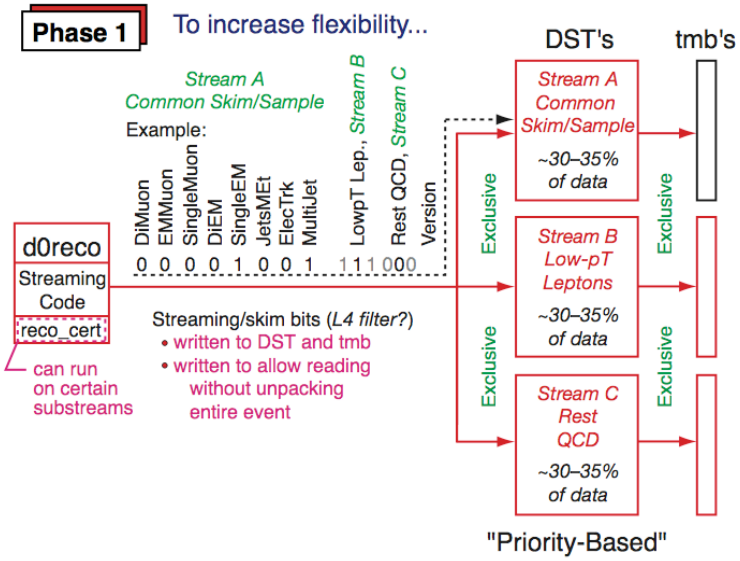

Figure 2: Detail of streaming/skim bits and an example of the "priority-based" sorting.

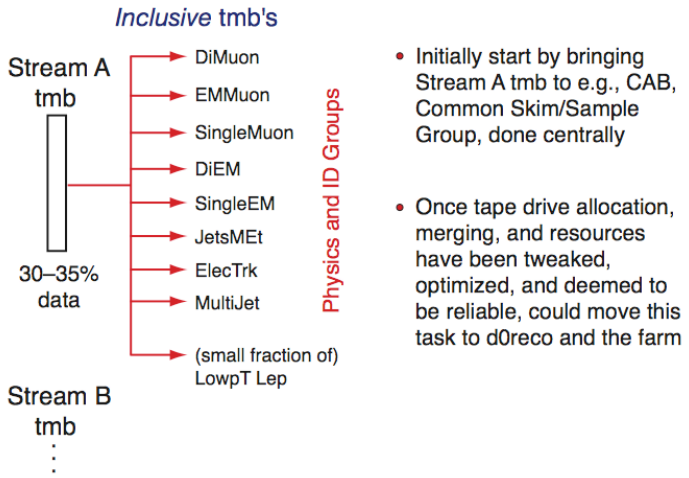

Figure 3: Example of further splitting of the Stream A (or any other stream) into inclusive thumbnails.

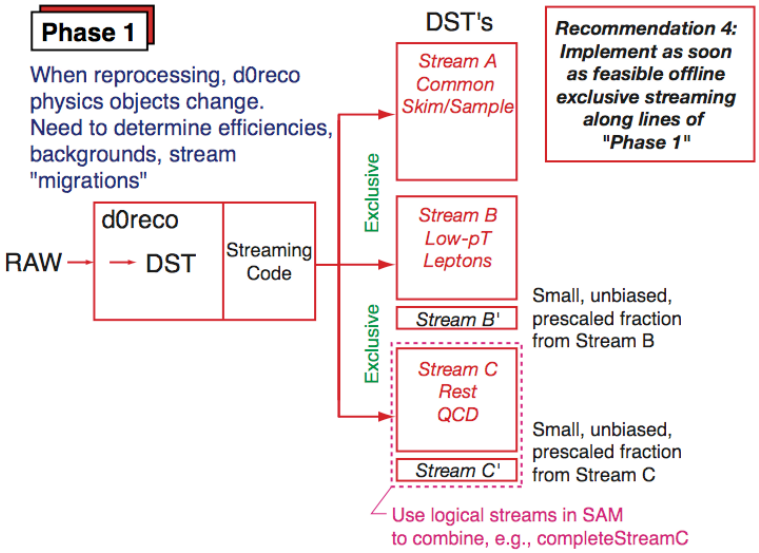

Figure 4: Further division in to small, unbiased prescaled fractions.